Multiple Disease Prediction Using Machine Learning

Dr.S.Gnanapriya[1]N.Anbarasan[2]

[1]Assistant Professor(SG),Department of Computer Application,Nehru college of Management Coimbatore,Tamilnadu,India.

[2] II MCA,Department of Computer Application,Nehru College of Management Coimbatore,Tamilnadu,India.

email:anbu47123@gmail.com

***Abstract*:**

Numerous machine learning approaches can do predictive analytics on vast volumes of data across a range of sectors. Although it is a challenging endeavor, predictive analytics in healthcare might ultimately help professionals make prompt judgments about patients' health and care based on vast amounts of data. Many people die from diseases including diabetes, breast cancer, and heart-related conditions worldwide, but the majority of these deaths are brought on by a failure to get regular checkups. A poor doctor-to-population ratio and a lack of medical infrastructure are the causes of the aforementioned issue. The data unequivocally demonstrate this; the WHO recommends a doctor-to-patient ratio of 1:1000, whereas India's doctor-to-population ratio is 1:1456, indicating a physician shortage. Diabetes, cancer, and heart disease can all pose a threat to humanity if they are not detected in time. Thus, many lives can be saved by early detection and diagnosis of these illnesses. The main goal of this effort is to use machine learning classification algorithms to anticipate dangerous diseases. Diabetes, heart disease, and breast cancer are all covered in this study. Our team created a medical test web application that uses machine learning to forecast various ailments in order to make this work smooth and accessible to the general audience. Our goal in this project is to create a web application that predicts numerous diseases, such as diabetes, heart disease, and breast cancer, using the idea of machine learning.

***Key words*:** Logistic Regression, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Diabetes, Breast cancer, Heart diseases.

1. INTRODUCTION: The leading causes of death in today's culture are mostly heart disease, diabetes, and breast cancer. The phrase "heart disease" describes a collection of conditions that impact your heart. Heart Disease (the faults of the heart you are born with) includes coronary artery disease, congenital heart defects, and arrhythmias (problems with heart rhythm).The phrase "heart disease" is frequently used in place of "cardiovascular disease**.**Convascular disease typically denotes a heart attack, angina (heart pain), or stroke. Heart diseases are problems that impact your heart's muscles or rhythm valves.[1] According to a Science Direct article, roughly 11 million people die in India each year, with cardiovascular disease accounting for 28% of those deaths.[2]In the US, a cardiovascular disease-related mortality occurs every 36 seconds, according to the "Centers for Disease Control and Prevention. . One type of cancer that arises in the breast tissue is called breast cancer. A change in the contour of the breast is one of the signs of breast cancer and also a lump in the breast, fluid eruption from the nipple, or reddish-

pink or scaly patches of skin. One of the highest mortality rates recorded across the globe was due to cancer. r. Cancer alone was the cause of 87 lakh deaths in 2015. After lung cancer, breast cancer has one of the highest reported fatality rates in terms of the number of deaths.[4]Around 2.4 lakh incidences of breast cancer are predicted to be the most common location in India by 2025, according to an article published by the Times of India on August 19, 2020. When your blood glucose level, commonly known as blood sugar, is extremely high, you may develop diabetes .The primary energy source that comes from the food you eat is blood sugar. The pancreas releases the hormone insulin, which helps to run metabolic processes by removing glucose from diet. According to the "International Diabetes Federation" across the globe, there are 42 lakhs of deaths caused by diabetes, and around 760 billion dollars USD are spent on diabetes (as a part of health expenditure).In India, over 10 lakh people die annually due to diabetes (Epidemiology of Diabetes),and according to the "Indian heart Association" nearly 11 crore individuals will end up suffering from diabetes by 2035. One illness is the focus of each inquiry in the previous AI models for medical care examination. For example, a diabetic inquiry, a cancer examination, a skin infection test, etc. There isn't a standard structure that allows one investigation to conduct many infection expectations. In our proposed system, we unify multiple diseases under a single user interface where you can perform predictions on Heart diseases, breast cancer, and diabetes. In this work, we are using the machine learning classification algorithms like Logistic Regression, Support Vector Machine (SVM), K- Nearest-Neighbors (KNN) to perform the prediction of multiple diseases.

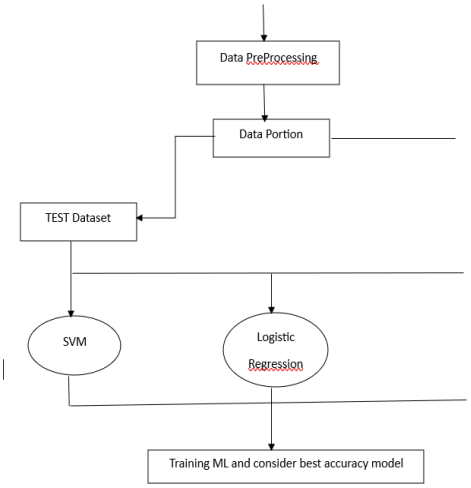
2. Relevent work: The analysis of earlier models for disease prediction that are relevant to our proposed work is covered in this part. Numerous investigations have been conducted to identify different illnesses. They haveapplied various data mining techniques for efficiently predicting a variety of diseases.[1]G NaveenKishore and few other authors proposed the work named Prediction Of Diabetes Using Machine Learning Classification Techniques proposed. This study applies a number of classification methods, including SVM, Logistic Regression, Decision Tree, KNN, and Random Forest, on 769 instances of the Pima dataset, which includes characteristics like body mass index, blood pressure, and pregnancy. The classification algorithm Random Forest has the best accuracy, recorded at 74.4%, while the KNN has the lowest accuracy, reported at 71.3%.[2] Gavin Pinto, Radhika Desai, and Sunil Jangid's work "Understanding the lifestyle of people to identify the reasons for Diabetes using data mining" covered diabetes sub-classification as well as lowering the risk of diabetes disease by data mining techniques. Using the information gathered through a Google Forms survey, the authors employed the Naïve Bayes and SVM classification algorithms. They claimed that the accuracy of SVM was 64.92 and that of Naïve Bayes was 60.44. [3]In the work presented by M.Marimuthu, S.DeivaRani, Gayatri. R described the cardio diseases in a detailed manner and also applied the
classification algorithms like SVM, Decision Tree, Naïve Bayes,K-Nearest Neighbors on the Framingham dataset from Kaggle.

In order to predict the risk of heart disease, the scientists compared a number of machine learning methods. The KNN classification algorithm in this work has the highest recorded accuracy of 83.60%. [4]Richa Sharma and Dr. Kanak Saxena address cardiovascular disease in the Purushottam-proposed work by utilizing Knowledge Extraction based on Evolutionary Learning, a Java programming technique for creating the development model for data mining problems. This work's greatest recorded accuracy is 86.7%.[5]M. Chinna Rao,
K. Ramesh, and G. Subbalakshmi presented a decision support system for heart disease prediction utilising the Nave Bayes classification method, which discussed the extraction of hidden information heart disease

dataset that can address complex queries.[13] A study by Amandeep Kaur and Jyothi Arora examined algorithms like KNN, SVM, ANN, and Decision Tree on the dataset of heart disease and plotted the accuracy graph. [14]Noreen Fatima proposed work on the Cancer forecast the data mining techniques and machine learning techniques thatcan predict cancer effectively on the large health records and described the study previous existing models.[15]Ch. Shravya, K.Pravallika, Shaik Subhani presented the work on Breast cancer prediction using Supervised machine learning techniques on the dataset and also analyzed the results with(PCA)principal component analysis and also used the dimensionality reduction and explained in a wellmannered way. [16]Nikitha Rane, Jean Sunny presented work on the classification of Cancer using machine learning concepts and their major discussion point is detecting cancer in very early stages so that a lot of lives can be saved.[17]Dilip Singh Sisodia ,Deepti Sisodia predicted diabetes using classification techniques and reported an accuracy of around 76% on the Pima dataset.[]d Using the KNN method,Dr. J. Ajayan, Dr. B. Santosh Kumar, and

T. Daniya have forecasted the occurrence of breast cancer with an accuracy rate of 83.33%.

3. Methodologies: The technique used in our suggested work is included in this section. As previously said, the goal of our study is to create a web application that uses machine learning models to identify conditions including diabetes, heart disease, and breast cancer. Themachine learning methodologies used in our proposed work are as follows:



3.1 Logistic Regression: The technique used in our suggested work is included in this section. As previously said, the goal of our study is to create a web application that uses machine learning models to identify conditions including diabetes, heart disease, and breast cancer. The following machine learning techniques are applied in our suggested work:Regression analysis using logistic The formula is Y=1/(1+EXPO(-value)). --(1) To anticipate the value of the output (referred to as y), input values (sometimes referred to as x) and co-efficients (Beta) are linearly mixed.

The equation for logistic regression is

y=EXPO(u0+u1*x)/(1+EXPO(u0+u1*x)). –

(2) y is the expected result, a0 is the bias or intercept, and a1 is the value of the only input and coefficient.

First-class (also known as default class) probability is predicted by logistic regression models. For instance, the default class may be male if we are developing a model to predict a person's gender based on height; this may be expressed formally as P(gender=male|height). Prediction probabilities must be converted to binary numbers, either 0 or 1, in order to be used.The logistic function is used to convert probabilities into predictions.One way to put the model together is as follows: y=EXPO(u0+u1*x)/(1+EXPO(u0+u1*x)). --(4)

After some computation, the equation is as in(p(x)/1-p(x))=u0+u1*x.--(5)

The odds of first-class or default class is the

name of the left-hand size equation (ratio). The probability of an occurrence is divided by the probability of its complement event to determine the odds. -ln(odds)=u0+u1*x With logistic regression, predictions are rather straightforward to execute. For illustration, suppose that a person's height of 150 can be used to determine their gender. Assuming that coefficients u0=-100 and y = EXPO(u0 + u1*X)
/ (1 + EXPO(u0 + u1*X)) for u1=0.6 EXPO(-100
+ 0.6*150) / (1 + EXPO(-100 + 0.6*150)) = -(7)
y y = 0.000045--(8)--(9) Males have a nearly negligible probability. We apply the particular probability in our constant practice.

3.2 K-Nearest Neighbors (KNN):

KNN is a machine learning method used for both classification and regression. Because it requires several iterations to get the highest accuracy, the approach is regarded as computationally costly. Because this method involves supervised machine learning, the data is labeled, and the computer gains the ability to anticipate the output derived from the data input. Additionally, even with enormous training data sets that contain noisy values, the algorithm performs flawlessly. The dataset is separated into test and training datasets by the algorithm. Model construction and training are done using the training dataset. The model is used to forecast the test data. We currently calculate the distance between the test point and the prepared knearest element values**.**

Distance Metrics: The most common distance measure used in KNN is the **Euclidean distance**, which is calculated as:

Euclidean Distance (2D): For two points P1(x1,y1)P_1(x_1, y_1)P1(x1,y1) and P2(x2,y2)P_2(x_2, y_2)P2(x2,y2), the Euclidean distance is given by:

d(P1,P2)=(x2−x1)2+(y2−y1)2d(P_1, P_2) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}d(P1,P2)=(x2−x1)2+(y2−y1)2

For higher dimensions, this formula generalizes as:

EuclideanDistance (N-Dimensional):For two points P1=(x1,x2,…,xn)P_1 = (x_1, x_2, \dots, x_n)P1=(x1,x2,…,xn) and P2=(y1,y2,…,yn)P_2 = (y_1, y_2, \dots, y_n)P2=(y1,y2,…,yn), the distance is:
d(P1,P2)=∑i=1n(xi−yi)2d(P_1, P_2) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}d(P1,P2)=i=1∑n(xi−yi)2

Where nnn is the number of features (dimensions).

How to choose a k value: The parameter denoted by K is the number of the closest neighbors. Finding the optimal k value to achieve the highest level of model precision is a challenging task. There isn't a predetermined, quantifiable method for figuring out the k value to achieve remarkable accuracy. The brute force method is the sole way to get a k value with remarkable precision, hence we must determine accuracy for a range of k values. The neighbor with the highest accuracy is taken into account for the forecast, along with the K values of neighbors 1 through 20.

3.3 Support vector machine:SVM is a method that requires supervision. Both regression and classification research make use of this approach. This technique employs coordinates to plot data in n-dimensional space.There are two forms of SVM: linear and nonlinear. We employ the linear SVM classifier in our study since the data we utilize is linearly separable. Finding the best hyperplane—which are the boundaries that separate the classes into categories—is how classes are classified. In two dimensions, the line is a hyperplane. In two dimensions, the line is adequate to divide the classes.Take the equation S0+(S1*U1)+(S2*U2)=0, for instance. The line's intercept is B2, and the coefficients are B0 and B1.The input points are K1 and K2.The categorization is done using this line.Above the line, the data value falls into the category since the value that the equation returns is greater than zero. "0."Below the line, the data point falls into category "1" since the value that the equation returns is less than zero.Classifying a point that produces a value around 0 is challenging. The gap between the line and the nearest data point is known as the margin.If the optimal line provides the greatest advantage, it can isolate the classes. as the

hyperplane of maximal margin. The perpendicular distance between the line and the closest highlight is used to register this margin. The data values are referred to as support vectors, and these points are essential for characterizing the line and building the classifier. Support vectors define and support hyperplanes.

3.4.1 Heart Disease Dataset: We have utilized UCI's "Heart Disease Dataset" to forecast the incidence of heart diseases. Thirteen medical predictor characteristics and one target feature make up this dataset. Chol, cp, trestbps, age, fbs, sex, restecg, exang, slope, thal, ca, oldpeak, and thalach are the characteristics. There are 75 attributes and 303 instances in the collection.

3.4.2 Diabetes Information Set : We used Kaggle's "Pima Indians Diabetes Dataset" to forecast the incidence of diabetic illnesses. Eight medical predictor features and one target feature are included in this dataset. The qualities are as follows: Pregnancy, bloodpressure, blood sugar, skin thickness, body mass index, insulin, age, and diabetes predisposition.

3.4.3 Breast Cancer Data Set: We have utilized Kaggle's "Breast Cancer Wisconsin (Diagnostic) Data Set" to forecast the incidence of breast cancer disorders. One target feature and thirty-one medical predictor features are included in this dataset. Among the crucial characteristics are the following: diagnosis, id, radius-mean, texture-mean, concavity-mean, concavity points-mean, smoothness-mean, area-mean, perimeter-mean, and area-mean.

4. Result: Our work has outperformed the accuracy listed in table 1. Using logistic regression for diabetes and breast cancer and KNN for heart disease, our work yielded the highest accuracy rates for these conditions, which are 77.60%, 83.84%, and 94.55%, respectively.

Disease	Logistic Regression	SVM	KNN	Best Accuracy
Heart Disease	82.50	78.87	83.84	83.84
Diabetes	77.60	75.64	75.52	77.60
Breast Cancer	94.55	91.38	92.55	94.55

Figure1 : Accuracy table

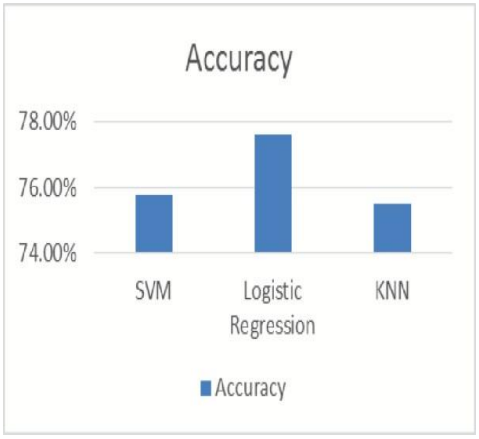


Figure2:accuracy values for diabetes**.**

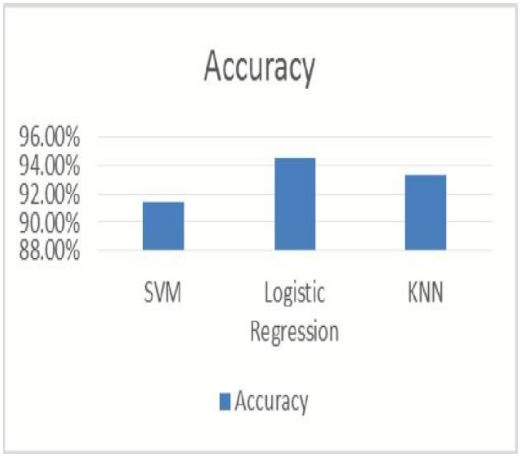


Figure3:accuracy values for Breast cancer

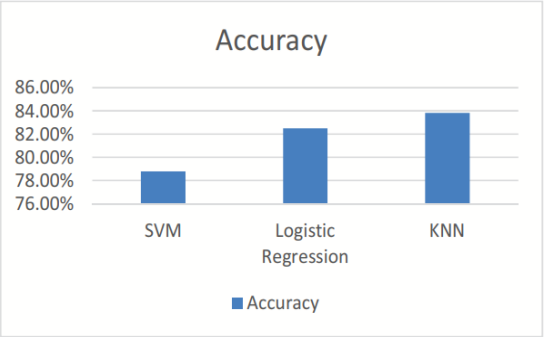


Figure4: curacy values for Heart Disease.

5. Conclusion: By employing the lightweight Flask API architecture to deploy the learned models, the proposed study unifies diabetes, heart disease, and breast cancer on a single platform. The models are trained using three classification algorithms; logistic regression and KNN provided good accuracy values for the disease prediction of diabetes and breast cancer, respectively, and KNN for the illness prediction of heart disease. The greatest value found from 1 to 21 neighbors is used to determine KNN's maximum accuracy. We can broaden our work in the future by including diseases that use deep learning models as well as those that are learned by machine learning models.

References:

[1] Trends in coronary Heart Disease Epidemiology

[2] Center for Disease Control and Prevention (Heart Disease Facts).

[3] Asian Pacific Journal of Global Trend of Cancer Mortality rate: A 25-year study.

[4] Times Of India: Cancer cases upswing 10% in 4 years to 13.9 lakh.

[5] International Diabetes Federation: Expenditure and deaths related to diabetes.

[6] Epidemiology of Diabetes :A report of Indian Heart Association.

[7] Naveen Kishore G,V .Rajesh ,A.Vamsi Akki Reddy, K.Sumedh,T.rajesh Sai Reddy, ”Prediction Of Diabetes Using Machine Learning Classification Algorithms”.

[8] Gavin Pinto, Sunil Jangid, Radhika Desai, ”Understanding the Lifestyle of people to identify the reasons of Diabetes using data mining”.

[9] M.Marimuthu ,S.Deivarani ,R.Gayatri, “Analysis of Heart Disease Prediction using Machine Learning Techniques”.

[10] Purushottam, Richa Sharma ,Dr. Kanak Saxena, ”Efficient Heart Disease Prediction System”.

[11] Adil Hussain She, Dr. Pawan Kumar Chaurasia,” A Review on Heart Disease Prediction using Machine Learning Techniques”.

[12] M. Chinna Rao ,K. Ramesh, G. Subbalakshmi,”Decision Support in Heart Disease Prediction System using Naïve Bayes”.

[13] Amandeep Kaur , Jyothi Arora,” Heart Disease Prediction using data mining Techniques :A survey”.

[14] Noreen Fatima , Li Liu , Sha Hong, Haroon Ahmed ,”Prediction of Breast Cancer, Comparitive Review Of Machine Learning Algorithms and their analysis”.

[15] Ch .Shravya ,K.Pravallika , Shaik Subhani, ”Prediction of Cancer using supervised machine learning Algorithms”.

[16] Nikita Rane, Jean Sunny, Rucha Kanade, Sulochana Devi,” Breast Cancer classification and prediction using machine learning “.

[17] Deepti Sisodia, Dilip Singh Sisodia,” Prediction of Diabetes using classification Techniques”.

[18] Dr.B.Santhosh Kumar, T.Daniya, Dr. J.Ajayan,” Breast Cancer Prediction using Machine Learning Algorithms”.

[19] Mümine KAYA KELEŞ ,”Cancer Prediction using and Detection using Machine Learning Algorithms : A Comparitive Study”.

[20] Heart Disease Dataset” by UCI.

[21] Pima Indians Diabetes Dataset” by Kaggle